International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

Transforming English Language Learning: Advanced Speech Recognition with MLP-LSTM for Personalized Education

M. Gangadharan

Assistant Professor of English, Kuppam Engineering College, Kuppam, Chittoor, Andhra Pradesh, India
Email: *gangadharanmgs[at]gmail.com*

Abstract: ***Speaking of speech recognitions within the English languages, it is the process of recognizing oral speeches and transcribings it into writing using exclusive algorithms. For the perishable skill of English language learning, use of innovative speech recognition technology using Advanced Speech Recognition Technologies MLP-LSTM is proposed in this paper to advance the existing online learning platforms. Previous research addresses the importance of NLP in English language learning but notes the challenges in effectively extracting and segmenting features from multimodal data. In order to overcome these problems, this paper incorporate the proposed MLP for feature extraction and LSTM for sequence learning. The utilization of MLP-LSTM provides not only a brilliant improvement of the capacity to transform spoken language and perceive it but also minimizes the Word Error Rate (WER) to 0.075. With this low WER, along with the total accuracy rate of 98.25 %, this paper focus on underlining how this system is more effective than traditional language learning tools. This paper has been implemented through Python Software. The given MLP-LSTM based speech recognition model lays the foundation for a highly complex yet accurate paced English language learning platform that will cater to the needs of the learners in the global scenario.***

Keywords: speech recognition, English language learning, MLP-LSTM model, Word Error Rate, Python software

1. Introduction

English language learning is the process of acquiring proficiency in English, which is the most widely spoken and studied language globally [1]. It includes strengthening speaking, listening, reading, and writing abilities, and in many areas, it is crucial for advancement in the class- room, the workplace, and personal life [2]. Learners come from a variety of language and cultural backgrounds and range in age from young children to adults [3]. Learning usually begins with fundamental abilities like syntax and vocabulary, then moves on to more intricate linguistic nuances and structures [4]. There are other approaches used, such as online courses, immersive learning, conventional classroom training, and language apps. Particularly successful are interactive and communicative methods that emphasise practical application and conversational practice [5]. The ability to learn English has been greatly improved by technological developments. Personalised and captivating experiences can be had using tools like speech recognition software, virtual classrooms, and language learning apps [6]. By offering opportunities for practice in real-world scenarios, adaptive learning pathways, and quick feedback, these technologies improve accessibility and efficiency of language acquisition [7]. Learning a language can be difficult due to factors including individual learning styles, motivation levels, and availability of high-quality materials. But as more and more digital tools and resources become available, learning outcomes continue to improve, making English competence accessible to a wider audience [8]. Speech recognition technology is transforming English language instruction by giving students individualised and engaging experiences [9]. With the use of this technology, computers can now understand spoken language and react accordingly, providing practice and real-time feedback- both of which are essential for language learning [10]. Tools for speech recognition provide many benefits for English language learners [11]. Without the assistance of a human teacher, they can improve their speaking abilities, practise pronouncing words correctly, and get instant feedback [12]. Speech recognition is used by pro-grammes like virtual assistants and language learning applications to generate dynamic and interesting lessons that are customised to the learner's skill level [13]. The accuracy and efficacy of these systems are improved by integrating sophisticated frameworks like LSTM networks and MLP [14]. Feedback from MLP-LSTM frameworks is accurate and context-aware since they are better at identifying different accents, intonations, and speech rates [15]. For non-native speakers who need to learn the subtleties of English, this is quite helpful. Speech recognition technology is becoming more and more reliable despite certain obstacles, such as handling background noise and various speech patterns [16]. Even though the technology developments hit the market, the current speech recognition systems using machine learning face challenges in handling diverse accents, dialects, and speech patterns [17]. They often require extensive training data and computational resources, and still struggle with background noise and context variability, leading to lower accuracy and reliability in real-world, multilingual environments [18]. So, this paper aims to enhance speech recognition for English language learning using MLP-LSTM, addressing current challenges like accent variability, noise interference, and contextual understanding for improved accuracy and personalization.

Motivated by the hybrid mechanisms in the recent technologies' development, this paper aims to leverage the advantages of combined MLP-LSTM framework for the detection of speech recognition for English language learning. Using the MLP and LSTM framework for speech recognition is a state-of-the-art method for processing and comprehending spoken language. This hybrid model incorporates the best aspects of both MLP for deep learning-

based pattern identification and LSTM for sequential data processing and long-term context maintenance [19]. The accuracy and performance of voice recognition are greatly improved by the MLP-LSTM framework. While LSTM networks capture temporal dependencies and context both essential for comprehending natural language, MLP processes complicated, non-linear interactions between speech components. By combining these two factors, the system can more effectively handle differences in speech speeds, accents, and background noise, leading to more accurate recognition results. This technology delivers revolutionary benefits for learning English. Students receive accurate and timely feedback on their grammar and pronunciation, allowing them to make corrections and increase their fluency instantly. The MLP-LSTM framework's adaptable feature enables customised learning experiences by adapting to the unique requirements and advancement of each student. The key contributions of this paper are as follows.

1) This study introduces a novel hybrid framework that combines Multilayer Perceptron (MLP) and Long Short-Term Memory (LSTM) networks. By leveraging the MLP's ability to handle high- dimensional data and the LSTM's capacity to manage long-term dependencies, the proposed framework aims to significantly enhance the accuracy and robustness of speech recognition systems.
2) The dataset used in this study is meticulously curated from Kaggle, consisting of a variety of voiced digit audio files from the Google Speech Commands Dataset and the Free Spoken Digit Dataset. The thorough pre-processing techniques, including noise reduction using the median filter method and normalization, ensure high-quality and consistent audio data for model training and evaluation.
3) The study employs Mel Frequency Cepstral Coefficients for feature extraction, a technique that effectively captures essential characteristics of audio signals. The detailed process of framing, windowing, Fourier Transform, power spectrum calculation, Mel-scale filter application, logarithm, and Discrete Cosine Transform (DCT) provides a robust feature set for subsequent analysis.
4) The proposed MLP-LSTM model is rigorously trained and evaluated using an 80:20 train-test split. The training process includes multiple layers of MLP with the Softmax activation function, dropout layers to prevent overfitting, and an LSTM layer to capture sequential dependencies. This structured approach ensures comprehensive learning and validation of the model.
5) The study presents a detailed algorithm outlining the steps for implementing the MLP-LSTM framework, from data loading and pre- processing to feature extraction, model building, training, evaluation, and prediction. This provides a clear roadmap for replicating and applying the proposed method in practical speech recognition applications.

The article's remaining section is organised as follows. The associated works are described in Section 2. The problem statement of the suggested paper is explained in Section 3. In Section 4, the MLP-LSTM network's construction was explained. Section 5 discusses the performance evaluation of the proposed MLP-LSTM network, and Section 6 wraps up the article proposed speech detection methods for studying English that use processing of natural languages research and extraction of characteristics. Word segmentation is a key function of NLP, a technique that aids computers in understanding human languages. This research introduces an advanced deep learning extraction of features technique for multipurpose characteristic retrieval using a multipurpose neural network for every phase. Features in distinct modes are converted to same-modal features by this method. The article presents a hybrid network method for segmenting keywords in English that takes into account the enduring connection among training time for forecasting approximation and textual interpretation. This technique solves long-distance dependency, reduces network training and prediction time, and annotates phrases in sequence using BI-GRU and the Conditional Random Field model. Investigations demonstrate that this approach delivers identical outcome results as BI-LSTM-CRF, but with a median predicted rate of processing which is 1.94 times speedier.

Weng, Qin, Tao, Pan, Liu, Li [23] proposed a speech recognition technique employed by deep learning enabled semantic communications. This study presents the creation of a deep learning-based conceptual interaction system for voice distribution, named Deep SC-ST. Voice generation and detection has been considered to be the interaction system's corresponding distribution roles. Following the extraction of speech recognition-related semantic information for distribution by a combined conceptual-channel encoder, the uttered phrase is recovered at the final location based on the obtained semantic characteristics. This greatly minimises the necessary amount of information to be trans- mitted while preserving efficiency. Subsequently, speech synthesis is carried out at the recipient, which is responsible for regenerating the voice outputs by providing a neural network modules with the recognised text and the speaker characteristics. The simulation findings demonstrate that the proposed DeepSC-ST operates far more effectively than the two the current DL-enabled communication devices and traditional systems for communication, especially in the low signal-to-noise ratio regime. A programme presentation is additionally built as an example of design for the DeepSC-ST. A limitation of the DeepSC-ST approach is its reliance on a high-quality conceptual-channel encoder, which may not perform optimally in highly variable or noisy environments.

Lin, Guo, Zhang [24], Chen suggested a single architecture for recognition of speech in several languages. An audio method, a phonological method, and a linguistic type are three fundamental elements that this study employs to merge linguistic speech recognition into a single structure. The primary objective is on reliable speech identification in air traffic control. The principal objective of this investigation is for the PM to convey the phoneme-based phrases that the AM transforms ATC voice into. Both phoneme- and word-based inaccuracies in the decoding results are corrected by the LM. It is suggested that a multiscale CNN architecture be used to suit the various data variations and enhance efficiency in order to handle radio trans- mission noise and speaker variability. A suggested machine translation PM with an encoder-decoder structure addresses phoneme-to-word conversion. By creating dependencies with frequent terms, RNN-based LMs are trained to take into account the code-switching peculiarity of the ATC speech. The ability to generalise the decoding performance is validated on multiple public corpora, and it is comparable to the end-to-end

model's, which makes it appropriate for real-time methods for assisting ATC tasks like security verification and ATC predictions. A limitation of this multilingual speech recognition architecture is its potential difficulty in handling highly specialized or rare phoneme variations not well-represented in the training data.

The papers present advancements in deep learning for speech recognition and NLP, including enhanced voice recognition algorithms, bilingual speech-to-text translation, efficient word segmentation, robust techniques, the study aims to achieve higher accuracy and robustness in converting spoken language into text. This approach is pivotal in advancing the capabilities of speech recognition technology, addressing challenges such as variability in speech patterns and environmental noise. The working methodology of this MLP-LSTM framework has been described in Fig. 1.

2. Problem Statement

The literature reveals significant advancements and challenges in integrating speech recognition technologies with deep learning for enhancing English language learning platforms. One study emphasizes the potential of deep learning in improving the accuracy of speech recognition systems by combining speech features and attributes. However, there is a need for more efficient algorithms that can seamlessly integrate these elements to enhance system performance further. Another study showcases the use of machine learning techniques to facilitate bilingual speech recognition, yet it highlights the complexities involved in achieving seamless language translation and voice recognition [20]. Additionally, research addresses the importance of NLP in English language learning but notes the challenges in effectively extracting and segmenting features from multimodal data. The issue of long-term dependency and prediction time also remains unresolved. Furthermore, a semantic communication system utilizing deep learning for efficient speech transmission has been proposed, but there is still a necessity for further optimization to handle varying signal-to-noise ratios. Lastly, a unified framework for multilingual speech recognition tailored for air traffic control has been presented, indicating a gap in generalizing such robust solutions for diverse and interactive language learning environments. These studies collectively underscore the need for innovative research to address these challenges and develop comprehensive, integrated solutions. In response to these gaps, the proposed study aims to revolutionize English language learning plat- forms by seamlessly integrating speech recognition technologies with deep learning, specifically utilizing a combination of MLP and LSTM networks. This approach seeks to leverage the strengths of MLP in handling high-dimensional data and LSTM in managing long-term dependencies and temporal patterns in speech, thereby addressing the aforementioned challenges and advancing the state-of-the-art in language learning technologies.

2.1 Data Collection

The dataset for this paper has been collected from Kaggle [25]. The data sets contain a variety of voiced digit audio files, which are crucial for building and evaluating voice recognition techniques. Three segments make up the Google Speech Commands Dataset: the audio folder has 10 recordings at 16 kHz for each digit (0–9), for a total of 100 re- cords; the validation folder has 100 recordings per digit, for a total of 1000 recordings at 16 kHz; and the test folder has 1000 recordings at 16 kHz for a thorough model evaluation. In addition, there are 160 recordings in the Free-Spoken Digit Dataset, which is housed in the free-spoken-digit-dataset folder. Each of the four speakers has four samples of each digit, all of which are recorded at 16 kHz. This dataset is perfect for training, validating, and testing speech recognition systems because of its diversity in speakers and recording situations. It ensures thorough performance assessment over a range of voices and digit pronunciations. While the dataset is publicly available, it is essential to respect the privacy and consent of individuals whose voices were recorded, ensuring that the data is used solely for research and development purposes. Additionally, proper acknowledgment of the data source and adherence to any licensing restrictions or terms of use pro- vided by Kaggle are crucial for ethical compliance.

It is the procedure of organising and preparing raw data so that it is ready for evaluation. Common steps in this technique include data purification, integration, transformation, reduction, discretization, handling missing values, copy removal, standardisation or normalisation of characteristics, and category-based variable encoding. The cleaning process consists of identifying and correcting mistakes and discrepancies in the data. By removing duplicates, each data point is guaranteed to be distinct. Improving or establishing characteristics ensures that each of the variables fall on an identical scale, which can enhance the efficiency of machine learning techniques. Effective data pre-processing ensures that the data is accurate, complete, and ready for further analysis, leading to stronger and more reliable model results. Pre-processing data is generally necessary to ensure its accuracy and consistency before evaluation. (Fig. 1)

2.2.1 Noise reduction by median filter method

The median filter significantly reduces noise in audio signals by substituting the average values across a given range for all samples of the audio. This technique removes impulsive noise while maintaining crucial aspects of speech signals, such as edges. Read the audio file, apply the median filter with a certain kernel size usually an odd number, like 3 or 5, then save the filtered output to put this into practice. By improving the audio data's quality, this pre-processing phase makes it better suited for speech recognition model testing, validation, and training. Unlike linear filters, the median filter preserves the edges and sharp transitions in speech signals, ensuring clarity in pronunciation and phonetic details essential for accurate speech recognition, especially in language learning applications. This makes it a suitable option over other filters, which may blur or distort important features of the speech signal. The better noise robustness and model performance can be guaranteed by using this filter consistently across all datasets training, validation, and test. The mathematical expression of the median method can be stated as follows

$$y[n] = median\ \{x\ [n — m], x\ [n — m + 1], \ldots, x\ [n + m — 1], x\ [n + m]\} \quad (1)$$

Here, $x[n]$ denotes the original signal
$y[n]$ denotes the filtered signal

$$m = \frac{k-1}{2}$$

k denotes the window size

2.2.2 Data normalization

In audio pre-processing, normalisation modifies the amplitude of audio signals to attain a uniform loudness throughout various recordings. With this method, distortion is avoided while scaling the signal to a desired peak amplitude, which for digital audio is usually between -1 and 1. The procedure entails determining the signal's maximum amplitude, computing a scaling factor, and then applying this factor to each and every sample in the signal. By supplying consistent and com- parable input levels, normalisation guarantees uniform volume levels, boosting audio quality and the efficiency of future audio processing activities, such speech recognition. The following is the mathematical representation of the normalisation process:

$$y[n] = \frac{x[n]}{A} \quad (2)$$

Here, $y[n]$ denotes the normalized signal
$x[n]$ denotes the original signal
A denotes the absolute maximum value of the signal

pass and the results between 0 and 1 regulate the gates in LSTM units. Tanh activation routines are also present to control the values entering and leaving the memory cell, which aids in maintaining the range of the variables. Below is a summary of the components.

The forget gate's primary job is to constantly determine which elements of the cell's state need to be retained and which ones need to be eliminated. This allows the LSTM system to focus on significant information and ignore irrelevant data through lengthy patterns thereby making it effective for issues needing the understanding of situations through extended periods of time, such as speech recognition, natural language processing, and time series forecasting. The forget gate assigns a value, ranging from 0 to 1, to each number in the cell state. A value close to 1 means "keep this information," whereas a value close to 0 means "forget this information."

$$f_t = \sigma\ W_f * [h_t - 1,\quad x_t] + b_f\ ) \cdots\cdots\cdots (3)$$

Where, f_t represents forget gate activation vector
σ represents sigmoid activation function
W_f represents weight matrix for the forget gate
h_{t-1} represents hidden state from the previous time step
x_t represents input at the current time step
b_f represents bias vector for the forget gate

Together with the forget gate, the input gate determines which extra information needs to be introduced to the cell state. The candidate cell state and gate activation are its two primary constituents. To enable the LSTM network to recognise and apply novel structures, the input gate's main function is to regulate the entry of new data to the cell state. The input gate assists the LSTM in striking the right balance among learning fresh, appropriate data from the input series and preserving essential long-term information by preferentially adjusting the cell state. By retaining appropriate data over time and continuously learning novel behaviours, this process enables LSTMs to perform activities involving a long-term contextual understanding, such as speech recognition, ma- chine translation, and time series forecasting.

$i_t = \sigma\ (W_i * [h_t -$

Where, i_t represents input gate activation vector W_i represents weight matrix for the input gate b_i represents bias vector for the input gate

The candidate values to be added to the cell state are computed as

$$\hat{C}t = \tanh\ (W_c\ [h_{t-1}, x_t] + b_c) \quad (4)$$

Where, $\hat{C}t$ represents candidate cell state vector
W_c represents weight matrix for the candidate cell state
b_c represents bias vector for the candidate cell state

The cell state, which serves as the network's memory and retains data over extended sequences, is a crucial part of LSTM networks. Because of its multi-time step design, it avoids the vanishing gradient issue that plagues conventional RNNs. Three gates alter the cell state: the forget gate, which chooses what data to remove; the input gate, which chooses what new data to add; and the output gate, which manages the cell state's output. Using this method, LSTMs may efficiently preserve and make use of long-term dependencies in sequential data.

$$C_t = f_t * C_t - 1 + i_t * \hat{C}t \quad (5)$$

Where, C_t represents updated cell state
C_{t-1} represents previous cell state

In an LSTM network, the information flow from the cell state to the hidden state is controlled at each time step by the output gate. Using the prior concealed state and the current input, it decides which portions of the cell state should be output. A sigmoid activation function, which generates values between 0 and 1, is in charge of the gate. The matching component of the cell state should be transmitted to the output if the value is close to 1, and it should be suppressed if it is close to 0. LSTMs can selectively expose information that is important for predictions or for layers further up in the network according to this technique.

$$O_t = \sigma\ (W_o\ [h_{t-1}, x_t] + b_o) \quad (6)$$

Where, O_t represents output gate activation vector
W_o represents weight matrix for the output gate b_o represents bias vector for the output gate

The hidden state is then computed as

$$h_t = O_t \times \tanh\ (C_t) \quad (7)$$

Where, h_t represents hidden state that is the output of the LSTM cell

Algorithm 1. MLP-LSTM Framework

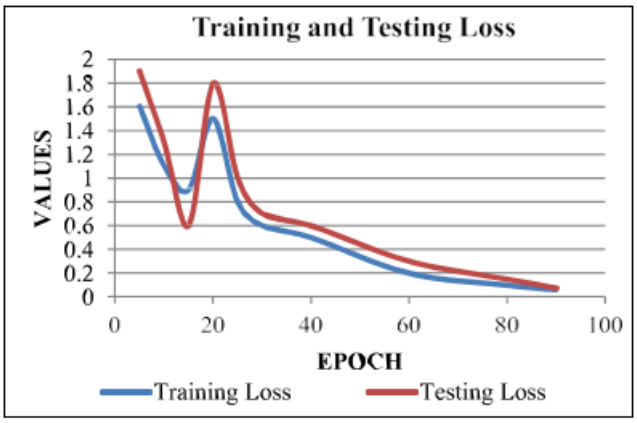

Figure 1: Training and Testing Loss.

Table 1: Comparison Values of Average Training Time with Existing Work

Model	*SQuAD*	*Wiki Text*	*Narrative QA*
BI — LSTM — CRF	182	290	142
BI — GRU — CRF	104	169	101
Proposed MLP — LSTM	91	103	85

Table 2: Comparison Values of Average Prediction Time with Existing Work.

Model	*SQuAD*	*Wiki Text*	*Narrative QA*
BI — LSTM — CRF[22]	189	289	141
BI — GRU — CRF[22]	96	148	92
Proposed MLP — LSTM	82	95	77

the workflow of an MLP-LSTM framework for speech recognition. It starts with the initialization and importing of necessary libraries. Audio files are then loaded and pre-processed to extract features like MFCC. The data is checked for validity and encoded into numerical labels. The model uses MLP layers for feature extraction and LSTM layers for sequence processing. After compiling, the model is trained, evaluated on test data, and used for predictions on new audio inputs. Error handling is incorporated throughout. (Figure 2)

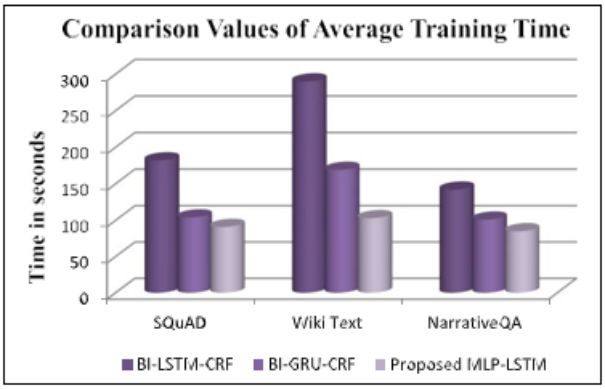

Figure 2: Comparison Values of Average Training Time

3. Results and Discussion

This section of the study on speech recognition using the MLP-LSTM framework presents findings on model performance and implications. This paper has been implemented by using Python software. It highlights achieved accuracy rates, training convergence, and generalization to new data through testing. Discussion focuses on the effectiveness of combining MLP for feature extraction and LSTM for sequence modelling, addressing challenges like noise robustness and variability in speech patterns. Insights into the framework's strengths and limitations are explored, suggesting future research directions for further enhancing speech recognition systems based on these findings. This section critically evaluates the methodology's success in meeting its objectives and contributing to the field.

3.1 Testing and training accuracy

Fig.3 visually represents the performance metrics of the developed model. It plots the accuracy scores obtained during the algorithm's implementation's training and testing phases. The training accuracy curve illustrates how well the model learns from the training data over successive epochs or iterations. Meanwhile, the testing accuracy curve indicates how effectively the model generalizes to unseen data, providing insights into its robustness and performance in real-world scenarios. Such charts are essential for evaluating and optimizing the MLP-LSTM framework to achieve high accuracy and reliability in speech recognition tasks.

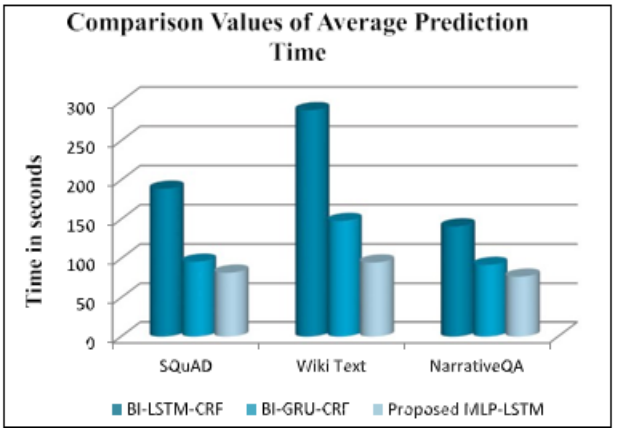

Figure 3: Comparison Values of Average Prediction Time.

3.2 Training and testing loss

Fig. 4 visualizes the progression of loss values throughout the model's training process. Loss measures how well the model predicts the target output compared to the actual output, with lower values indicating better performance. The training loss curve depicts how quickly the model converges during training, ideally decreasing over epochs as the model learns. Conversely, the testing loss curve evaluates the model's ability to generalize to new data, showing trends in performance on unseen samples. Monitoring these charts aids in optimizing the MLP-LSTM framework for accurate and efficient speech recognition applications.

3.3 Calculation of average training time

The average training time for a single sentence in speech recognition, model complexity, computational resources, and implementation efficiency. Typically, the training time T can be expressed as follows.

$$T = L \times C \times F \quad (8)$$

Where L represents the length of the sentence in seconds
C represents the computational complexity per second of

audio.
F represents factor accounting for hardware performance and parallelization efficiency

Table 1 compares the average training times (in seconds) for single sentences across three datasets: SQuAD, Wiki Text, and NarrativeQA, using three models. The BI-LSTM-CRF model exhibits the longest training times, with 182 s for SQuAD, 290 s for Wiki Text, and 142 s for NarrativeQA, reflecting the computational demands of LSTMs and CRFs. The BI-GRU-CRF model is more efficient, with reduced times of 104, 169, and 101 s due to the faster GRU cells. The proposed MLP-LSTM model achieves the shortest training times, at 91, 103, and 85 s, show- casing its optimized architecture and training efficiency across all datasets. Fig. 6 describes the comparison values of average training times for single sentences with existing works.

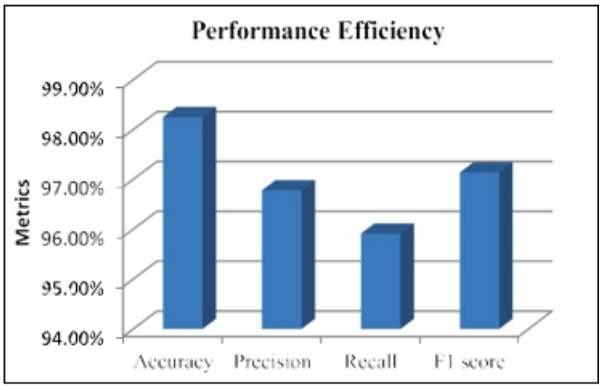

Figure 4: Performance Metrics for the Speech Recognition

3.4 Calculation of average prediction time

The average prediction time for a speech-recognition sentence in- volves processing the input audio through the trained model. Typically, prediction time is faster than training, often taking milliseconds to a few seconds per sentence on modern hardware, depending on model complexity and optimization. The average prediction time can be calculated using the following formula.

$$T_{pre} = L \times C1_{pre} \times P$$

Where, *L* represents the length of the sentence in seconds
C_{pre} represents the computational complexity per second of audio
P represents the processing power

Table 2 compares the average prediction times (in seconds) for single sentences across three datasets: SQuAD, WikiText, and NarrativeQA, using different models. The BI-LSTM-CRF model is the slowest, with times of 189 s, 289 s, and 141 s for SQuAD, WikiText, and NarrativeQA, respectively, due to the complexity of bidirectional LSTMs and CRF layers. The BI-GRU-CRF model improves efficiency, reducing times to 96 s, 148 s, and 92 s for the respective datasets, leveraging the faster GRU cells. The proposed MLP-LSTM model is the fastest, achieving times of 82 s, 95 s, and 77 s, demonstrating superior optimization and performance for rapid prediction. Fig. 7 describes the comparison values of average prediction times for single sentences with existing works.

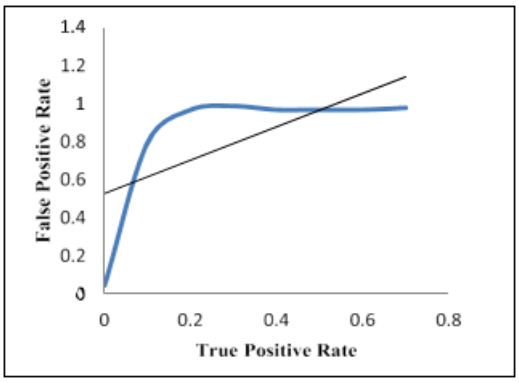

Figure 5: ROC Curve

3.5 Performance Metrics of Proposed Method

Performance measures measure how well algorithms translate spoken language into text. These metrics assess the system's accuracy in understanding and transcribing speech while accounting for variables such as speaker unpredictability, accents, and background noise. Metrics measuring memory utilization and processing speed are also used to evaluate the system's viability. Speech recognition technologies, virtual assistant programs, transcription services, and accessibility tools can all be improved with better performance and dependability by carefully analyzing these parameters. Fig. 8 depicts the performance efficiency of this research.

3.5.1 Word Recognition Accuracy

Word recognition accuracy quantifies the performance of a speech recognition system by measuring the percentage of correctly transcribed words. It is calculated by comparing the system's output to a reference transcript, accounting for errors such as substitutions, deletions, and insertions. WRA reflects the system's performance, with higher per- centages indicating better accuracy. The formula is:

$$WRA = 1 - \left(\frac{S + D + I}{N}\right) \times 100\% \qquad (9)$$

Where, *S* represents the number of substitutions
D represents the number of deletions
I represents the number of insertions
N represents the total number of words

3.5.2 Word error rate

Word Error Rate is a common metric used to evaluate the accuracy of speech recognition systems. It measures the proportion of words incorrectly predicted by the system. A lower WER indicates better performance. For instance, a WER of 10 % means that 10 % of the words were incorrectly recognized, reflecting the system's error rate in transcription. The formula for WER is:

$$WER = \frac{S + D + I}{N} \times 100\% \qquad (10)$$

Where, *S* represents the number of substitutions
D represents the number of deletions
I represents the number of insertions
N represents the total number of words

3.5.3 Word correct rate

Word Correct Rate is a metric used to evaluate the accuracy of a speech recognition system by measuring the proportion of correctly recognized words out of the total words in the reference transcript. WCR indicates the percentage of words correctly recognized without considering insertions. A higher WCR signifies better performance. The for

$$WCR = N - S - D \times 100\% \quad (11)$$

Where, *S* represents the number of substitutions
D represents the number of deletions
N represents the total number of words

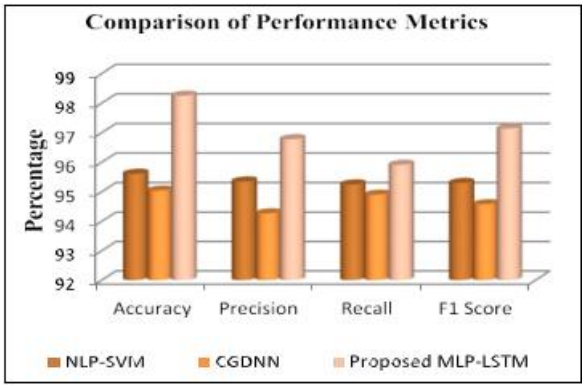

Figure 6: Comparison of Performance Metrics Chart

3.5.4 Accuracy

The system's capacity to accurately and error-free translate spoken words into text is called accuracy. Metrics like Word Error Rate (WER) and Character Error Rate (CER), which quantify differences between the spoken words and the recognized text, are commonly used to measure it. High accuracy denotes fewer transcription errors, indicating a better comprehension of various dialects, speech patterns, and contextual factors. This study refines algorithms utilizing linguistic models and training data to maximize accuracy. This improves the use of the algorithms in voice-controlled systems, transcription services, and language translation. Accuracy measures are essential to assess developments and guarantee consistent performance in voice recognition technology.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (12)$$

The percentage of words that are accurately transcribed relative to all words that the system has determined to be correct is known as precision. Its main goal is to reduce false positives- recognizing erroneous words as correct. Because precision minimizes errors and pre- serves fidelity to the spoken input, it is essential to ensure the transcribed text's accuracy. Advanced algorithms, linguistic models, and context-aware processing methods can all lead to higher precision. To improve usability and user satisfaction with more dependable outputs, this statistic is crucial for optimizing voice recognition systems in various applications, including virtual assistants, dictation software, and accessibility aids.

$$Precision = \frac{TP}{(TP + FP)} \quad (13)$$

3.5.5 Recall

Recall is the percentage of accurately transcribed words relative to all words that ought to have been recognized as correct. It strongly emphasizes reducing false negatives- words that are mispronounced or overlooked. High recall guarantees thorough coverage of spoken input, precisely catching all pertinent terms. Optimizing recall requires fine-tuning algorithms, combining various linguistic models, and improving context awareness. This measure is essential for enhancing the accuracy and completeness of transcriptions in voice-activated systems, automated transcription services, and language learning applications, among other uses. The reliability and usability of voice recognition systems can be improved in various real-world circumstances by optimizing recall.
Recall TP= (TP + FN)

3.5.6 F1 score

The F1 score in voice recognition papers is a composite score that measures total accuracy by combining recall and precision parameters. It balances these measures to offer a thorough assessment of transcribing performance. The F1 score provides a reliable evaluation of the system's accuracy in transcribing spoken language while reducing false positives and negatives. It is computed as the harmonic mean of precision and recall. By improving algorithms, modifying thresholds, and adding language models, F1-Score was maximized. This statistic is essential for evaluating speech recognition systems across various languages, dialects, and ambient circumstances to ensure dependable performance in real-world applications.

$$F1score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \quad (14)$$

3.5.7 ROC curve

Fig. 9 shows the ROC curve, in which the area under the curve quantifies the overall ability of the model to discriminate between positive and negative classes. A model with an AUC of 1.0 indicates perfect classification, while an AUC of 0.5 suggests performance no better than random chance. ROC curves are useful for comparing different models and selecting the optimal threshold for classification decisions.

Table 3 contains the comparison values of the Word Error Rate of existing works and the proposed method. The table compares the WER of different speech recognition models. The CNN-HMM model has a WER of 0.185, while the DNN-HMM improves to 0.105 using deep neural networks. The CNN-RBM model, combining CNNs and restricted Boltzmann machines, achieves a WER of 0.095. Adding an adaptive sequence attention technique in CNN-RBM-ASAT further reduces WER to 0.082. The proposed MLP-LSTM model, utilizing multi-layer perceptrons and long short-term memory networks, performs the best with a WER of 0.075, indicating the highest accuracy in recognizing speech.

3.6 Discussion

During this paper's training and testing phase, several important discoveries were discovered. A broad dataset has been used to improve model robustness, including a range of speech styles, accents, and environmental conditions during training. Effective learning of temporal relationships in

speech sequences, a critical skill for accurate transcription, was demonstrated using the MLP-LSTM architecture [27,28]. The model's promising Word Error Rate (WER) of 0.075 % during testing suggests that it can accurately transcribe speech. Evaluations of precision and recall scores also demonstrated the system's capacity to reduce false positives and false negatives in transcription tasks [20,22, 26] Computational efficiency was also prioritized, and the MLP-LSTM framework showed respectable processing speeds appropriate for real-time applications. It could be better to innovate and improve these technologies to support language learners and educators in achieving proficiency and fluency in spoken English, even though this study rep- resents a significant advancement in using MLP-LSTM for speech recognition in English language learning.

4. Conclusion

In summary, this work on MLP-LSTM framework voice recognition has shown encouraging outcomes and possible directions for further research. With a competitive Word Error Rate (WER) and strong performance across a range of accents and speaking styles, the MLP-LSTM architecture demonstrated its efficacy in capturing temporal relation- ships and nuances in spoken English. This technology, which offers precise and customized feedback on pronunciation and spoken fluency, has significant implications for improving language learning experiences. Future research in this field may concentrate on several important areas. Initially, adding more accents and linguistic variances to the dataset would improve the generalization and flexibility of the model. Furthermore, adding real-time feedback systems to instructional mate- rials might provide students with prompt remediation recommendations, speeding up their language learning. Investigating cutting-edge methods like transformer structures or attention processes could help further optimize the MLP-LSTM framework and increase computing efficiency and transcription accuracy. Furthermore, adding multimodal inputs like speaking mixed with gestures or facial expressions could enhance the educational process and offer more thorough feedback on communication abilities. Resolving the model outputs' interpretability and guaranteeing error analysis openness are also essential for building user confidence and improving instructional applications. Lastly, longitudinal research could evaluate how speech recognition technology affects language competency and student involvement over the long run. In short, future research projects will continue to innovate and improve these technologies to better support language learners and educators in achieving proficiency and fluency in spoken English, even though this study represents a significant advancement in using MLP-LSTM for speech recognition in English language learning. Future research could benefit from interdisciplinary collaborations with cognitive science to enhance understanding of language acquisition processes and with human-computer interaction experts to refine user interfaces and engagement strategies. Integrating these fields could lead to more effective and intuitive speech recognition systems for language learners.

Acknowledgement
The authors extend their appreciation to the Deanship of Scientific Research at the Northern Border University, Arar, KSA for funding this research work through the project number "NBU-FPEJ-2024–1584-02".

References

[1] P. Ma, S. Petridis, M. Pantic, Visual Speech Recognition for Multiple Languages in the Wild, Nat. Mach. Intell. vol. 4 (11) (2022) 930–939, https://doi.org/10.1038/s42256-022-00550-z.
[2] A.S. Subramanian, C. Weng, S. Watanabe, M. Yu, and D. Yu, "Deep Learning based Multi-Source Localization with Source Splitting and its Effectiveness in Multi-Talker Speech Recognition," Nov. 28, 2021, arXiv: arXiv:2102.07955. Accessed: Jun. 18, 2024. [Online]. Available: ⟨http://arxiv.org/abs/2102.07955⟩.
[3] Baevski, W.-N. Hsu, A. Conneau, and M. Auli, "Unsupervised Speech Recognition," 2022.
[4] L. Wu, L. Wu, Research on Business English Translation Framework Based on Speech Recognition and Wireless Communication, Mob. Inf. Syst. (2021).
[5] Y. Fathullah et al., "Prompting Large Language Models with Speech Recognition Abilities," Jul. 21, 2023, arXiv: arXiv:2307.11795. Accessed: Jun. 18, 2024. [Online]. Available: ⟨http://arxiv.org/abs/2307.11795⟩.
[6] J. Wu, Deep spiking neural networks for large vocabulary automatic speech recognition, Front. Neurosci. vol. 14 (2020).
[7] M. Ravanelli et al., "Multi-task self-supervised learning for Robust Speech Recognition," Apr. 17, 2020, arXiv: arXiv:2001.09239. Accessed: Jun. 18, 2024. [Online]. Available: ⟨http://arxiv.org/abs/2001.09239⟩.
[8] J.-Y. Hsu, Y.-J. Chen, and H. Lee, "Meta Learning for End-to-End Low-Resource Speech Recognition," Oct. 26, 2019, arXiv: arXiv:1910.12094. Accessed: Jun. 18, 2024. [Online]. Available: ⟨ http://arxiv.org/abs/1910.12094⟩.
[9] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix, and T. Nakatani, "Far-Field Automatic Speech Recognition," Sep. 20, 2020, arXiv: arXiv: 2009.09395. Accessed: Jun. 18, 2024. [Online]. Available: ⟨http://arxiv. org/abs/2009.09395⟩.
[10] C.-C. Chiu, J. Qin, Y. Zhang, J. Yu, and Y. Wu, "Self-Supervised Learning with Random-Projection Quantizer for Speech Recognition," 2022.